

Selection by Partitioning the Solution Paths

Yang Liu

Fred Hutchinson Cancer Research Center

and

Peng Wang *

Department of Operations, Business Analytics and Information Systems
University of Cincinnati

June 22, 2016

Abstract

The performances of penalized likelihood approaches profoundly depend on the selection of the tuning parameter; however there has not been a common agreement on the criterion for choosing the tuning parameter. Moreover, penalized likelihood estimation based on a single value of the tuning parameter would suffer from several drawbacks. This article introduces a novel approach for feature selection based on the whole solution paths rather than choosing one single tuning parameter, which significantly improves the selection accuracy. Moreover, it allows for feature selection using ridge or other strictly convex penalties. The key idea is to classify the variables as relevant or irrelevant at each tuning parameter and then select all the variables which have been classified as relevant at least once. We establish the theoretical properties of the method, which requires significantly weaker conditions compared to existing literature. We also illustrate the advantages of the proposed approach with simulation studies and a data example.

Keywords: tuning parameter, variable selection, solution paths, penalized likelihood, selection criteria.

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

1 Introduction

The penalized likelihood approach has been rather popular in the past decade for feature selection problems, where it is of interests to select a subset of relevant covariates. In order to achieve the purpose, one would need to compute the solution paths and then choose a tuning parameter with some criterion. The solution paths yielded with the chosen tuning parameter are considered as the estimates of the parameters.

Consider the following linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (1)$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$ is an n -dimensional response vector, $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^T$ is a p -dimensional vector of regression coefficients, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ is an $n \times p$ design matrix, and $\boldsymbol{\varepsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ is an n -dimensional vector of i.i.d. random errors. Without loss of generality, we can assume in model (1), $\mathbf{x}_j, j = 1, \dots, p$ are standardized and the response \mathbf{y} are centered, i.e., $\sum_{i=1}^n x_{ij} = 0, \sum_{i=1}^n x_{ij}^2 = n, j = 1, \dots, p, \sum_{i=1}^n y_i = 0$.

For linear regression problems described in (1), the penalized likelihood approach is equivalent to the penalized least squares (PLS) regression, where the coefficients are estimated by minimizing the following objective function:

$$\frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^p J(|\beta_j|), \quad (2)$$

where $J(\cdot)$ is a penalty function that controls the number of nonzero coefficients, and $\lambda > 0$ is a tuning parameter. Unlike traditional variable selection procedures, the penalized least squares approach can carry out variable selection and estimation simultaneously, since the objective function (2) automatically shrinks estimates of some coefficients to zeros.

Furthermore, theoretical properties of the penalized least squares estimators are established when the tuning parameter is appropriately chosen (Fan and Li, 2001; Fan and Peng, 2004; Zhao and Yu, 2006; Zou, 2006; Zhang, 2010; Bühlmann and van de Geer, 2011).

Apparently, the estimator of β , denoted by $\hat{\beta}(\lambda) = \{\hat{\beta}_1(\lambda), \dots, \hat{\beta}_p(\lambda)\}^T$, is a function of the tuning parameter λ once the penalty function $J(\cdot)$ is specified. In practice, one can compute $\hat{\beta}(\lambda)$ for a number of different values of λ to obtain the solution paths of all the coefficients and then choose a tuning parameter λ using some type of criterion. The purpose of the tuning criterion is to find a λ that balances the fit and the complexity of the model. Therefore, one needs to specify both a penalty function and a criterion to select the tuning parameter λ in order to carry out variable selection with the penalized regression approach.

Much research has been devoted to the development of the penalty function. In general, there are two classes of penalty functions, convex penalties and non-convex penalties. The L_1 penalty, referred to as the lasso (Tibshirani, 1996), is probably the most commonly used convex penalty. Zou (2006) proposed the adaptive lasso approach which corrects the bias of the lasso for nonzero regression coefficients by adding a weight to the L_1 penalty. As for the non-convex penalties, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty by using a quadratic spline function with knots at λ and $a\lambda$, where $a > 2$ is a constant; Zhang et al. (2010) proposed the minimax concave penalty (MCP) by minimizing the maximum concavity of the model for variable selection and unbiasedness; Shen et al. (2012) also proposed the truncated L_1 penalty (TLP) as a surrogate of the L_0 penalty. These non-convex penalties all enjoy the oracle property in the sense that estimators obtained by applying these penalties are as efficient as if the nonzero coefficients are already known.

Another aspect of the feature selection involves the selection of the tuning parameter.

Some general selection criteria include cross validation, generalized cross validation, AIC, BIC, GIC. Chen and Chen (2008) pointed out that these criteria usually identify too many irrelevant features when the number of variables is large. Such phenomenon has also been described in Broman and Speed (2002), Siegmund (2004) and Bogdan et al. (2004) in their studies of quantitative loci mapping. Chen and Chen (2008) proposed the extended BIC (EBIC), which promotes model sparsity by adjusting BIC with an additional penalty term for the growing number of parameters in the model. Recently, Sun et al. (2013) also proposed a new technique via variable selection stability, which directly focuses on the selection of the informative variables.

Although the above criteria have been well studied for more than a decade, there has been no concurrence of opinion on which criterion to employ for the choice of the tuning parameter. See, for examples, Table 1 for a list of publications on major statistics and machine learning journals and the different criteria they use. In fact, the currently used feature selection procedure, using only one chosen value for the tuning parameter, may suffer from inevitable drawbacks that it is often impossible to correctly identify all the features, no matter which criterion we use. In the following, we demonstrate these drawbacks with a simulated example.

We would use the following simulated example to illustrate the aforementioned problem. Suppose there are 10 nonzero (relevant) variables and 30 zero (irrelevant) ones in the model (1), where the coefficients of these 10 nonzero variables are $\beta_1^* = \dots = \beta_5^* = 3, \beta_6^* = \dots = \beta_{10}^* = -2$. The entries of the variables $\mathbf{x}_j, j = 1, \dots, p$ are generated from the standard normal distribution. The pairwise correlation between the first 10 variables is 0.9. The remaining 30 variables are independent with each other, and are also independent with the first 10 variables. Furthermore, we generate the error from the normal distribution $N(0, 3^2)$ and we set the sample size to be $n = 50$. This example is proposed by Wang et al. (2011),

which is designed to study the performance of the existing variable selection methods for the data with complicated correlation structure.

We apply the R package *lassoshooting* to obtain a set of parameter estimators $\hat{\beta}(\lambda)$ at each value of λ and plot the lasso solution paths in Figure 1. We pick the grid of the tuning parameters on the log scale, therefore we use the log values of the tuning parameter as the x -axis in the plot. The dashed lines in Figure 1 represent the solution paths for nonzero variables and the solid lines represent those for the zero ones. The tuning parameters chosen by the 2-fold CV, GCV, AIC, BIC and the extended BIC (EBIC) are shown by the vertical lines. We report the total number of the selected variables, the number of false positives (FP, the number of selected zero variables) and the number of false negatives (FN, the number of missed nonzero variables) by these criteria in Table 2. Here the true model is known, therefore we also record the result of the “oracle” selection in a sense that we select the best possible tuning parameter which minimizes the number of incorrect selections (i.e., the number of selected zero variables + the number of missed nonzero variables).

We observe that CV, GCV, AIC, and BIC tend to select too many spurious variables and the extended BIC tends to drop most of the nonzero variables. Even for the “oracle” selection, many nonzero variables are excluded in the model. The problem looks more evident when we focus on three lines in the lower panel of Figure 1. Here the two dotted lines (1 and 3) are the solution paths of two nonzero coefficients, while the solid line 2 is the solution path for an zero one. Apparently selecting a small λ , as AIC BIC and GCV do, misleads us to identifying all the three coefficients as nonzero. On the other hand, a large λ , as CV, EBIC and Oracle select, incorrectly shrinks both coefficients of the nonzero variables to zero. As a matter of fact, it is impossible to correctly identify all the three features regardless of the value of the tuning parameter we choose, although one can even tell the differences between the three features by simply observing the solution paths.

We observe that CV, GCV, AIC, and BIC tend to select too many spurious variables and the extended BIC tends to drop most of the nonzero variables. Even for the “oracle” selection, many nonzero variables are excluded in the model. The problem looks more evident when we focus on three lines in the lower panel of Figure 1. Here the two dotted lines (1 and 3) are the solution paths of two nonzero coefficients, while the solid line 2 is the solution path for an zero one. Apparently selecting a small λ , as AIC BIC and GCV do, misleads us to identifying all the three coefficients as nonzero. On the other hand, a large λ , as CV, EBIC and Oracle select, incorrectly shrinks both coefficients of the nonzero variables to zero. As a matter of fact, it is impossible to correctly identify all the three features regardless of the value of the tuning parameter we choose, although one can even tell the differences between the three features by simply observing the solution paths.

The above restriction of utilizing just one tuning parameter could seriously reduce the accuracy of the feature selection in general, since solution paths like this in Figure 1 happen quite often no matter which penalty we employ. This is especially true when there exist large correlations among the variables or the dimensions of the features are extremely high. To overcome this restriction, we develop an innovative and intuitive approach, which utilizes the whole solution paths to improve the selection accuracy. Our approach can correctly identify the relevance of the features like those 3 ones (the labeled lines 1, 2, and 3) in Figure 1.

We achieve the objective by developing a partitioning rule that cuts the whole solution paths into two regions, namely “zero region” and “nonzero region”. First, we develop a new clustering method which divides all the variables into two clusters, the relevant set and the irrelevant set, for each value of the tuning parameter λ . Then the whole plot of the solution paths can be partitioned into two regions by the red curves as shown in Figure 2. We name the region inside the two red curves as the zero region, and that outside

as the nonzero region. Finally, we choose all the variables, which have been identified as a relevant variable for at least one value of λ , as the important features. We consider a feature unimportant if its solution path never goes out of the zero region. We name the above procedure *selection by partitioning the solution paths (SPSP)*. It can be well observed from Figure 2 that this SPSP procedure correctly selects 9 out of the 10 relevant variables and drops all irrelevant ones, outperforming the result from any single value of λ in terms of selection accuracy. Another advantage of the SPSP is that it does not require the coefficients of the unimportant variables shrunk to zero and it allows us to carry out feature selection with just a ridge regression.

We consider a feature important even if its solution path enters the nonzero region just once. The strategy may seem aggressive in identifying relevant variables. This is because, we start the SPSP process rather conservative, in the sense that for the smallest value of λ , we consider every variable “unimportant”. We initiate the partitioning process with the smallest λ . Then the clustering at a larger value of λ depends on the results from the previous λ . Therefore, the SPSP procedure combines a conservative starting point with an aggressive selection strategy to optimize the selection accuracy.

The SPSP procedure is connected with the stability selection approach, proposed in a seminal discussion paper by Meinshausen and Bühlmann (2010). Their approach is based on the probabilities of the variables being selected, and these probabilities are obtained from generic sub-sampling approach. Therefore the stability selection does not require the selection of the tuning parameter and also successfully utilizes the information of the whole solution paths. However, the SPSP procedure would work on any shrinkage penalty function, while with stability selection, one would still need to employ a penalty that can shrink coefficients to zero. Moreover, the computational cost of the SPSP procedure is much smaller as no sub-sampling is involved and we only need to compute the solution

paths once. In addition, the cut off probability in stability selection is arbitrary, while for SPSP, the constant playing the similar role is data-adaptive. Finally, we find from simulation studies that the stability selection tends to select too few variables, therefore produces a higher false negative rate compared to the SPSP.

This work is also remotely related to Bayesian variable selection approaches, where the tuning parameters or candidate models are assigned a prior distribution, and the posterior distributions of the models are evaluated. A related idea of applying a collection of models for variable selection is the Bayesian model averaging approach, which calculates the posterior probability that a variable enters the model by averaging over all of the models Hoeting et al. (1999), Raftery et al. (1997), Posada and Buckley (2004). Another similar idea is from Barbieri and Berger (2004), which constructs the posterior inclusion probabilities for all the features using Bayesian model averaging technique. The final model would include all the variables whose posterior probabilities of being in the model are 0.5 or higher. The so-called probability median model also has the flavor of utilizing the results from different tuning parameters, rather than just choosing one of them.

The article is organized as follows. Section 2 introduces the SPSP approach. Section 3 discusses the consistency and oracle properties of the SPSP estimator. Section 4 presents the results from simulation examples and Section 5 provides an application in biology to detect the significant genes for glioblastomas.

2 Selection by Partitioning the Solution Paths

In this section, we propose an approach which utilizes the whole solution paths to select the informative features in the model. Firstly, we develop a partition rule to divide the variables into two groups: relevant and irrelevant, at each tuning parameter. Based on the

partitioning rule, we propose the main contribution of the paper: *selection by partitioning the whole solution paths (SPSP)*. The proposed approach allows us to improve the selection accuracy by efficiently combining all the information across the whole solution paths, especially in cases when strong correlations among the variables are presented.

Considering penalized least squares problem in (2), we denote the index set for the true relevant (nonzero) variables is $S = \{j : \beta_j^* \neq 0\}$ with $s = |S|$, and the index set for irrelevant variables is $S^c = \{j : \beta_j^* = 0\}$. The goal of variable selection is to correctly recover this sparsity pattern from the noisy observations in the model, and correctly estimate S .

Once we specify the penalty function $J(\cdot)$ in (2), a grid of the tuning parameters is required to compute the solution paths. Typically, we would pick the grid to be equidistant on the log scale as follows: $\lambda_{\min} = \lambda_1 < \dots < \lambda_K = \lambda_{\max}$, where $\lambda_{\min} = 1/n$, λ_{\max} is the smallest λ yielding $\hat{\beta} = 0$. Bühlmann and van de Geer (2011), Shen et al. (2012) also suggested the same way to build the grid of the tuning parameters. Note that since the solution paths are usually continuous with respect to the tuning parameter, they vary little for the choice of the grid as long as enough tuning parameters are selected.

For each λ_k , we obtain a vector of the penalized least squares estimators as $\hat{\beta}_k = (\hat{\beta}_{k,1}, \dots, \hat{\beta}_{k,p})^T$. A variable is more likely to be identified as relevant if its estimator is farther away from 0, regardless of the sign of the estimator. Therefore we take the absolute values of the estimators as $\hat{\beta}_k^{(abs)} = (|\hat{\beta}_{k,1}|, \dots, |\hat{\beta}_{k,p}|)^T$.

In general, variables with a larger $|\hat{\beta}_{k,p}|$ are more likely to be important. Therefore, we are interested in finding a proper cutoff point $T_k = T(\lambda_k)$, such that the estimated relevant set \hat{S}_k and irrelevant set \hat{S}_k^c at $\lambda = \lambda_k$ are derived as

$$\hat{S}_k = \{j : |\hat{\beta}_{k,j}| > T_k\} \quad (3)$$

and

$$\hat{S}_k^c = \{j : |\hat{\beta}_{k,j}| \leq T_k\}. \quad (4)$$

To obtain $(T_k, \hat{S}_k, \hat{S}_k^c)$ for each λ_k , we sort the absolute values $|\hat{\beta}_{k,1}|, \dots, |\hat{\beta}_{k,p}|$ in ascending order to obtain $\hat{\beta}_{k,(1)}^{(abs)} \leq \dots \leq \hat{\beta}_{k,(p)}^{(abs)}$, where $\hat{\beta}_{k,(j)}^{(abs)}$ is the j th order statistics of $|\hat{\beta}_{k,1}|, \dots, |\hat{\beta}_{k,p}|$. Then we define the adjacent distances between these ordered values as

$$D_{k,j} = \hat{\beta}_{k,(j)}^{(abs)} - \hat{\beta}_{k,(j-1)}^{(abs)}, j = 1, \dots, p.$$

Note that $D_{k,1}$ is the adjacent distance between $\hat{\beta}_{k,(1)}^{(abs)}$ and 0 as we define $\hat{\beta}_{k,(0)}^{(abs)} = 0$ for convenience. Let $\hat{s}_k = |\hat{S}_k|$ be the number of variables in the estimated relevant set \hat{S}_k , then there are $p - \hat{s}_k$ variables in the estimated irrelevant set \hat{S}_k^c . Hereafter, by (3) and (4), we simply define the gap between \hat{S}_k and \hat{S}_k^c as the adjacent distance between $\hat{\beta}_{k,(p-\hat{s}_k)}^{(abs)}$ and $\hat{\beta}_{k,(p-\hat{s}_k+1)}^{(abs)}$, i.e.,

$$D(\hat{S}_k, \hat{S}_k^c) = D_{k,p-\hat{s}_k+1} = \hat{\beta}_{k,(p-\hat{s}_k+1)}^{(abs)} - \hat{\beta}_{k,(p-\hat{s}_k)}^{(abs)}.$$

In principle, $D(\hat{S}_k, \hat{S}_k^c)$, the gap between \hat{S}_k and \hat{S}_k^c , should be sufficiently large to separate the irrelevant features from the important ones. We consider $D(\hat{S}_k, \hat{S}_k^c)$ large enough if it meets the following two criteria,

$$\frac{D_{\max}(\hat{S}_k)}{D(\hat{S}_k, \hat{S}_k^c)} \leq R, \quad (5)$$

$$\frac{D(\hat{S}_k, \hat{S}_k^c)}{D_{\max}(\hat{S}_k^c)} > R, \quad (6)$$

where $D_{\max}(\hat{S}_k) = \max\{D_{k,j} : j > p - \hat{s}_k + 1\}$ is the largest adjacent distance in \hat{S}_k , $D_{\max}(\hat{S}_k^c) = \max\{D_{k,j} : j < p - \hat{s}_k + 1\}$ is the largest adjacent distance in \hat{S}_k^c and R is a certain constant. The criterion (5) ensures that the gap between the relevant set and

irrelevant set should have the same order as the distances between the estimated nonzero coefficients, while (6) guarantees that $D(\hat{S}_k, \hat{S}_k^c)$ has a higher order than the distances between the estimators of the zero coefficients. The constant R is used to control the differences of the magnitudes between the estimators of the zero coefficients and those of the nonzero coefficients.

The principles (5) and (6) here are equivalent to the claim that the order of the estimators for the nonzero coefficients should be the same, and it should be higher than those of the zero coefficients. Instead of comparing every pair of the estimators, we just use adjacent distances for simplicity of the calculation. Therefore, finding the proper T_k now transforms to finding an adjacent distance that is large enough—satisfies (5) and (6)—to be the gap between the estimators for the zero coefficients and those for the nonzero ones.

In order to introduce our proposed algorithm for partitioning the solution paths, we further define the largest adjacent distance under where $D_{\max}(\hat{S}_k^c)$ happens in \hat{S}_k^c as

$$D_{\max 2}(\hat{S}_k^c) = \max\{D_{k,j} : j < j', D_{k,j'} = D_{\max}(\hat{S}_k^c)\}.$$

Then following the aforementioned principles, we develop the algorithm for partitioning the solution paths as follows.

Selection by Partitioning the Solution Paths (SPSP)

- 1 Set the initial values as $T_0 = \infty$, $\hat{S}_0 = \emptyset$, $\hat{S}_0^c = \{1, \dots, p\}$, and proceed to λ_1 .
- 2 At each λ_k , we estimate $T_k, \hat{S}_k, \hat{S}_k^c$ from $T_{k-1}, \hat{S}_{k-1}, \hat{S}_{k-1}^c$ and $\hat{\beta}_k^{(abs)}$.
 - 2.1 Update $T_k = \max_{j \in \hat{S}_{k-1}^c} |\hat{\beta}_{k,j}|$, $\hat{S}_k = \{j : |\hat{\beta}_{k,j}| > T_k\}$, $\hat{S}_k^c = \{j : |\hat{\beta}_{k,j}| \leq T_k\}$;
 - 2.2 Calculate $D_{k,1}, \dots, D_{k,p}$. Further obtain $D_{\max}(\hat{S}_k^c)$, $D_{\max 2}(\hat{S}_k^c)$ and $D(\hat{S}_k, \hat{S}_k^c)$.

2.3 If $D(\hat{S}_k, \hat{S}_k^c) \leq R \times D_{\max}(\hat{S}_k^c)$ and $D_{\max}(\hat{S}_k^c) > R \times D_{\max 2}(\hat{S}_k^c)$, we update

$$T_k = \hat{\beta}_{k,(j'-1)}^{(abs)}, \hat{S}_k = \{j : |\hat{\beta}_{k,j}| > T_k\}, \hat{S}_k^c = \{j : |\hat{\beta}_{k,j}| \leq T_k\}.$$

Otherwise $T_k, \hat{S}_k, \hat{S}_k^c$ remain unchanged as in Step 2.1.

3 Proceed to λ_{k+1} and repeat Step 2 until $k = K$.

4 Identify the union of all \hat{S}_k as the index set for our selected relevant variables, i.e.,

$$\hat{S} = \bigcup_{k=1}^K \hat{S}_k.$$

At each λ_k , we find in Step 2 the cutoff point T_k , the location of the gap that distinguishes the relevant and irrelevant variables, based on the results from λ_{k-1} . This not only simplifies the computation process, but also makes the boundary line $T_k = T(\lambda_k)$ relatively more smooth to avoid unstable selection results. Specifically, in Step 2.1, we first use the largest estimated coefficients among those identified as “zero-coefficients” for λ_{k-1} as the current boundary. This could take care of the case where some coefficients in \hat{S}_{k-1} becomes small and enters into the zero region at λ_k . At Step 2.2 and Step 2.3, we decided on whether any adjacent distances within \hat{S}_k^c is large enough to be considered as the new gap between the zero and nonzero coefficients. This manages the scenario that there are too few variables in \hat{S}_k so that a “large” gap still exists.

For λ_1 , we use the initial values set in Step 1, where all the variables are considered “irrelevant”, which means we are conservative in identifying the relevant variables at the start of this process. This is because we implement an aggressive selection strategy at Step 4 to use the union of all \hat{S}_k ’s as our estimated index set for the relevant variables, allowing us to minimize the false positive rate at each λ_k . Another reason is that the estimation at the small λ_k ’s are usually unstable, because the design matrix corresponding to nonzero

coefficients is usually ill-conditioned. Therefore it is better to select fewer relevant variables, rather than taking the high risks of committing false positive errors.

Here the choice of the constant R for the partitioning rule is data-adaptive. In practice, we first obtain the estimator with a small value of λ , then take the absolute value and compute the adjacent distances of the sorted values. We then choose the constant R as the ratio of the maximal adjacent distance to the second maximal adjacent distance. Simulation studies confirm that the strategy is practically effective. In fact, both the simulations and our theoretical results show that our final results are not sensitive to the choice of R .

Once we identify the index set of all the relevant variables \hat{S} , we estimate the regression parameters $\hat{\beta}_{\hat{S}}$ a model that only includes the features that has been selected, $\mathbf{y} = \mathbf{X}_{\hat{S}}\boldsymbol{\beta}_{\hat{S}} + \boldsymbol{\varepsilon}$, where $\mathbf{X}_{\hat{S}} = (\mathbf{x}_j)_{j \in \hat{S}}$, $\boldsymbol{\beta}_{\hat{S}} = (\beta_j)_{j \in \hat{S}}^T$. In most cases, the number of features in \hat{S} is smaller than the sample size, we just use the least squares estimator as $\hat{\beta}_{\hat{S}}$. If the number of selected variables are larger than the sample size, we could use a ridge regression with a small shrinkage factor.

It can be seen that one advantage of this procedure is that it can not only be applied to penalties like lasso, adaptive lasso, SCAD, MCP, but also can it be applied for the penalties which do not produce the sparse solutions, such as the ridge penalty. Therefore, it could greatly reduce computation complexity for feature selection problems, since strictly convex penalties like ridge are easier to solve.

In addition, the SPSP algorithm can be easily extended to handle selection problems in a wide range of models including graphical modeling, generalized linear models, Cox's proportional hazards models. In these models, we usually apply the penalized likelihood approach, which obtains a sparse estimate by solving an objective function consisting of likelihood and a penalty function. As a result, we can apply the SPSP algorithm on penalized likelihood estimators with a similar fashion. A simulation example using SPSP

in graphical models is provided in Section 4.

3 Consistent Feature Selection

In this section, we discuss the advantages and properties of SPSP procedure with lasso for feature selection on high dimensional data ($p \gg n$). Here, we limit our efforts to linear regression, although the SPSP procedure is generally applicable to other selection problems as well. The technical proofs of all the lemmas and theorems in the section are put in a separate supplementary file.

Consistent variable selection for a procedure refers to the following property of its estimator \hat{S}

$$P(\hat{S} = S) \rightarrow 1, \text{ as } n \rightarrow \infty.$$

In most existing literature, it is only possible to achieve feature selection consistency if the tuning parameter is restricted to a specific intervals. Moreover, the widths of these intervals are usually so small that they converge to 0, see, for example, Fan and Li (2001), Fan and Peng (2004), Zhao and Yu (2006) and Zou (2006). The advantage of our SPSP procedure is that it circumvent the issue by utilizing the whole solution path, instead of trying to choose a proper tuning parameter, which is notoriously difficult.

It is well-known that under the high dimensional setting ($p > n$ or $p \gg n$), the Gram matrix $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ is degenerate, which raises many difficulties in controlling the values of the lasso estimator. Therefore, some conditions on the design matrix are always required to establish the consistency of feature selection. The most typical condition is probably the following irrepresentable condition (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Zou, 2006; Yuan and Lin, 2007): $\left| \frac{1}{n}\mathbf{X}_{S^c}^T\mathbf{X}_S \left(\frac{1}{n}\mathbf{X}_S^T\mathbf{X}_S \right)^{-1} \right| \leq 1 - \eta$, where η is a positive constant. Zhao and Yu (2006) showed that the above condition is sufficient and

almost necessary for lasso to be selection consistent. However, the condition is restrictive and difficult to verify in practice. Here, we will first show that the SPSP procedure is selection consistent under a much weaker compatibility condition (Bühlmann and van de Geer, 2011) or the restricted eigenvalue condition (Bickel et al., 2009) if we could bound the tuning parameter to an interval decided by the maximum adjacent distance of β^* . In order to achieve selection consistency over all values of tuning parameter, we need another identifiability condition, which is still weaker compared to the irrerepresentable condition.

We first introduce the following compatibility condition as Assumption 1.

Assumption 1. *Compatibility Condition* (van de Geer, 2007; Bühlmann and van de Geer, 2011) *For some constant $\phi > 0$, and for any vector $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1$, the following compatibility condition holds:*

$$\|\boldsymbol{\delta}_S\|_1^2 \leq \left(\boldsymbol{\delta}^T \hat{\Sigma} \boldsymbol{\delta} \right) s / \phi^2.$$

Here the compatibility condition is built on the fact that the lasso bias $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*$ satisfies $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1$ with probability close to 1 (Bühlmann and van de Geer, 2011; Bickel et al., 2009). Hence we can restrict ourselves to such vectors in the condition. Similarly, several related assumptions have also been proposed to establish the consistency property of the lasso, such as the restricted eigenvalue condition (Bickel et al., 2009), the restricted isometry condition (Candes and Tao, 2005), and the coherence condition (Bunea et al., 2007). The further relations among these conditions can be found in Bühlmann and van de Geer (2011).

Lemma 1. *Suppose the compatibility condition holds. Let $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log p}{n}}$ for any $t > 0$,*

then for $\lambda \geq 2\lambda_0$, with probability at least $1 - 2e^{-t^2/2}$, we have

$$\frac{1}{n} \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{4\lambda^2 s}{\phi^2}.$$

This lemma implies the bound for the prediction error and the following bound for l_1 -error of the lasso estimator:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{4\lambda s}{\phi^2}.$$

The compatibility condition required to bound the above errors is substantially weaker than the irrepresentable condition, which is necessary for achieving consistent variable selection under the currently used framework of choosing a single λ . Bühlmann and van de Geer (2011) showed that the irrepresentable condition actually implies the compatibility condition.

With the proposed SPSP procedure, we are able to accomplish consistent variable selection without the irrepresentable condition. Because at each λ , we cluster the lasso estimators into two groups, rather than labeling all the variables with non-zero estimates as important features, we need only to bound the bias of the lasso estimators, rather than shrinking some coefficients to zeros. The following theorem shows that when λ is not too large, the SPSP procedure identifies the true relevant set S with probability close to 1 with just the compatibility condition.

Let $\delta_\lambda = \frac{4\lambda s}{\phi^2}$ and $\delta_0 = \delta_{\lambda_0}$, where λ_0 is defined in Lemma 1. We first sort the absolute values of true non-zero coefficients in ascending order to get $|\beta^*|_{(1)}, \dots, |\beta^*|_{(s)}$, and define the true adjacent distance as $D_0 = 0, D_1 = |\beta^*|_{(1)}, D_2 = |\beta^*|_{(2)} - |\beta^*|_{(1)}, \dots, D_s = |\beta^*|_{(s)} - |\beta^*|_{(s-1)}$. Let $C_0 = \sqrt{\frac{D_{\max} + \delta_0}{\delta_0}} - 1$, $D_{\max} = \max_{1 \leq i \leq p} \{D_i\}$ and

$$C = \frac{D_{\max}}{\min\{|\beta_i^*| : |\beta_i^*| > (2 + C_0)\delta_0\}}.$$

Moreover, let

$$C_{under}^i = \frac{D_i}{\max\{D_{i'} : i' < i\}},$$

for $i = 2, \dots, s$ and $C_{under}^1 = \infty$, and

$$C_{upper}^i = \frac{D_i}{\max\{D_{i'} : i' > i\}},$$

for $i = 1, \dots, s-1$ and $C_{upper}^s = \infty$. Further define $\hat{D}(\Theta)$ as the largest adjacent distance of $\hat{\beta}_i$ for all $i \in \Theta$, where $\hat{\beta}_i$ are obtained with tuning parameter λ , i.e. $\hat{D}(\Theta) = \max_{j \in \Theta} \min_{j' \in \Theta} \left| |\hat{\beta}_j| - |\hat{\beta}_{j'}| \right|$, and $\hat{D}(\Theta_1, \Theta_2) = \min_{j \in \Theta_1, j' \in \Theta_2} \left| |\hat{\beta}_j| - |\hat{\beta}_{j'}| \right|$. In addition, let $R = 1 + C$.

Theorem 1. *Let $i_\lambda = \min\{i : C_{under}^i \geq R, C_{upper}^i \geq \frac{1}{C}, D_i > (1 - \frac{R}{C_{under}^i})^{-1}(1 + R)\delta_\lambda\}$ and $S_\lambda = \{j : |\beta_j^*| \geq |\beta_{i_\lambda}^*|\}$. Under the compatibility condition, if $\lambda > 2\lambda_0$, the following inequalities hold for lasso estimator with probability at least $1 - 2e^{-t^2/2}$,*

$$\begin{aligned} \frac{\hat{D}(S_\lambda, S_\lambda^c)}{\hat{D}(S_\lambda^c)} &> R, \\ \frac{\hat{D}(S_\lambda)}{\hat{D}(S_\lambda, S_\lambda^c)} &\leq R. \end{aligned}$$

Denote the important features identified at λ as \hat{S}_λ . Theorem 1 implies that $P(\hat{S}_\lambda = S_\lambda) > 1 - 2e^{-t^2}$. When the tuning parameter λ are bounded by $\min_{j \in S} |\beta_j^*| > (1 + R)4\lambda s/\phi^2$, we recover S exactly with probability close to 1,

$$P(\hat{S}_\lambda = S) > 1 - 2e^{-t^2}.$$

Further we conclude that for larger λ , i.e. $\max_{j \in S} |\beta_j^*| > (1 + R)4\lambda s/\phi^2 \geq \min_{j \in S} |\beta_j^*|$,

there are no false positive signals in \hat{S}_λ , $P(\hat{S}_\lambda \subset S) > 1 - 2e^{-t^2}$. The following theorem then follows immediately from the fact that our SPSP estimator is $\hat{S} = \cup_\lambda \hat{S}_\lambda$.

Theorem 2. *Let $i_{2\lambda_0} = \min\{i : C_{under}^i \geq R, C_{upper}^i \geq \frac{1}{C}, D_i > (1 - \frac{R}{C_{under}^i})^{-1}(1 + R)2\delta_{\lambda_0}\}$ and $S_{2\lambda_0} = \{j : |\beta_j^*| \geq |\beta_{(i_{2\lambda_0})}^*|\}$. Under the compatibility condition, the SPSP procedure \hat{S} over $\lambda \in [2\lambda_0, \frac{\phi^2 D_{\max}}{4s(1+R)})$ recovers $S_{2\lambda_0}$ with probability at least $1 - 2e^{-t^2}$,*

$$P(\hat{S} = S_{2\lambda_0}) > 1 - 2e^{-t^2}.$$

In particular, when $\min_{j \in S} |\beta_j^| > (1 + R)2\delta_0$,*

$$P(\hat{S} = S) > 1 - 2e^{-t^2}.$$

Theorem 2 suggests that the proposed SPSP procedure is consistent for variable selection under just the compatibility condition. With Theorem 2, we require that the tuning parameter λ is not larger than $\frac{\phi^2 D_{\max}}{4s(1+R)}$, the lower bound of which could be obtained with prior information. When no such information is available, we would need that the SPSP estimator over larger λ will not select any noise variables. This is easy to verify in practice, since we can just take a look at whether any new variables enter into the relevant set for larger values of λ . However, in order to theoretically guarantee such behavior of the solution paths, we need an additional condition, which is still substantially weaker than the irrepresentable condition.

Assumption 2. Identifiability Condition *Let $\eta > 0$ be some constant, for any possible lasso solutions $\hat{\beta} = (\hat{\beta}_S, \hat{\beta}_{S^c})$, the following identifiability condition holds:*

$$\|\mathbf{X}\beta^* - \mathbf{X}_S \hat{\beta}_S - \mathbf{X}_{S^c} \hat{\beta}_{S^c}\|^2 \geq \min_{\|\beta_S\|_1 \leq \|\hat{\beta}_S\|_1 + (1-\eta)\|\hat{\beta}_{S^c}\|_1} \|\mathbf{X}\beta^* - \mathbf{X}_S \beta_S\|^2. \quad (7)$$

The above identifiability condition indicates that with the true set of relevant variables, we can approximate the noiseless response $\mathbf{X}\boldsymbol{\beta}^*$ at least as well as with any other set of variables under any constraint of l_1 norm. It is not hard to verify that the condition is weaker than the irrepresentable condition.

Lemma 2. *The irrepresentable condition implies the identifiability condition.*

In fact, we can further weaken the identifiability condition by allowing $\boldsymbol{\beta}_{S^c}$ to be non-zero on the right side of (7). Instead, we only require $\|\boldsymbol{\beta}_{S^c}\|_1$ to be smaller than $\|\boldsymbol{\beta}_S\|_1$ up to a constant k . Moreover, we relax the inequality (7) by $\kappa\eta\|\hat{\boldsymbol{\beta}}_{S^c}\|_1$ to obtain the following weak identifiability condition.

Assumption 3. Weak Identifiability Condition *Let $\eta > 0$ be some constant, for any possible lasso solutions $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\beta}}_{S^c})$, the following identifiability condition holds for some κ :*

$$\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\hat{\boldsymbol{\beta}}_S - \mathbf{X}_{S^c}\hat{\boldsymbol{\beta}}_{S^c}\|^2 \geq \min_{\boldsymbol{\beta} \in \Theta(\|\hat{\boldsymbol{\beta}}_S\|_1, \|\hat{\boldsymbol{\beta}}_{S^c}\|_1)} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\boldsymbol{\beta}\|^2 - \kappa\eta\|\hat{\boldsymbol{\beta}}_{S^c}\|_1,$$

where $\Theta(\|\hat{\boldsymbol{\beta}}_S\|_1, \|\hat{\boldsymbol{\beta}}_{S^c}\|_1) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leq \|\hat{\boldsymbol{\beta}}_S\|_1 + (1 - \eta)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1, \|\boldsymbol{\beta}_{S^c}\|_1 \leq k\|\boldsymbol{\beta}_S\|_1\}$.

From now on, we refer the above weak identifiability condition with constants k and κ as $WIC(k, \kappa)$. The condition is to ensure that when λ is large, the lasso estimates for the zero coefficients would not be much larger than those of the non-zero coefficients, so that we would not have any false positive signals from our SPSP procedure \hat{S} . We combine that consideration with Theorem 1 to obtain the following result for the whole solution paths under the compatibility condition and $WIC(k, \kappa)$.

Theorem 3. *Under the compatibility condition and $WIC(k, \kappa)$ with $k = \frac{2}{2s+Rs(s+1)}$, suppose*

$$D_{\max} > \lambda_0 \frac{4s(1+R)}{\phi^2} \left\{ \frac{Rs^2 + (2+R)S + 2}{\eta} - 1 + \kappa \right\},$$

then the SPSP procedure over $\lambda \in [2\lambda_0, \infty)$ identifies $S_{2\lambda_0} = \{j : |\beta_j^*| > (1 + R)2\delta_0\}$ with probability at least $1 - 2e^{-t^2}$, i.e.

$$P(\hat{S} = S_{2\lambda_0}) > 1 - 2e^{-t^2}.$$

When the true values of the coefficients are of higher order than $\sqrt{\frac{\log p}{n}}$, it follows immediately from Theorem 3 that the asymptotic probability of identifying the true relevant set S is 1. Note we that we only need the tuning parameter λ here not to be too small— $\lambda > 2\lambda_0$; we do not require the tuning parameter λ to be in a specific region, which is typically required to obtain similar results in existing literature.

Theorem 3 is derived under conditions weaker than the irrepresentable condition, because with the SPSP procedure, we do not require consistent variable selection for any value of λ . We only need to control the bias of the lasso estimators for smaller λ , and control the l_1 norm of the lasso estimators of those zero coefficients for larger λ , both of which are weaker results compared to achieving selection consistency at certain value of λ . By combining these weaker results, the SPSP procedure is able to accomplish feature selection consistency without proper choice of tuning parameter under substantially weaker conditions.

4 Simulation Studies

We mainly propose the SPSP algorithm for the variable selection problems with high correlations in the data sets, especially for the high dimensional data problems. Therefore, we present several simulations with relatively complicated correlation structures here. The first one illustrates one low dimensional case, where the relevant variables are highly cor-

related. For the remaining high dimensional simulations, the relevant variables are highly correlated with different signs for the second one. The third model examines the proposed methods for sparse models with only few informative covariates among a large number of redundant ones. The last simulation is a misspecified model where the true model is not contained in the model space.

All the following simulations are generated from the linear model (1) with $x_{ij} \sim N(0, 1)$, $i = 1, \dots, n$, $j = 1, \dots, p$ and $\epsilon_i \sim N(0, \sigma^2)$. The details of the simulation setups are described as follows.

- (M1) Let $\beta^* = (3, 3, 3, 3, 3, -2, -2, -2, -2, -2, 0, \dots, 0)$, where the first 10 coefficients are nonzero and the remaining 30 coefficients are zero. The pairwise correlation between the first 10 variables is 0.9. The remaining 30 variables are independent with each other, and are also independent with the first 10 variables. We set $n = 50$ and $\sigma = 3$.
- (M2) Let $\beta^* = (3, 3, -2, 3, 3, -2, 0, \dots, 0)$, where the first 6 coefficients are nonzero and the remaining 94 coefficients are zero. The pairwise correlation between the first 3 variables is 0.9, the pairwise correlation between the second 3 variables is also 0.9 and the remaining 94 variables are independent with each other. Furthermore, the first 3 variables, the second 3 variables, and the remaining 94 variables are independent with each other. We set $n = 50$ and consider two different noise levels $\sigma = 1$ and 3.
- (M3) Let $\beta_1^* = 3, \beta_2^* = 1.5, \beta_5^* = 2$, and the remaining coefficients are zero. The correlation between x_{j_1} and x_{j_2} is $0.5^{|j_1 - j_2|}$. We set $n = 50$, $\sigma = 2$ and two different dimensions $p = 100$ and $p = 1000$.
- (M4) The true coefficients in the model is $\beta^* = (1, -1.25, 0.75, -0.95, 1.5, 0, \dots, 0)$, where among these 100 coefficients, the first 5 are nonzero and the remaining 95 are zero.

All the variables are independent of each other, i.e, the rows of \mathbf{X} is generated from $N(\mathbf{0}, \mathbf{I}_p)$. We consider $n = 50$ and $\sigma = 0.25$ in the model. Here the response y is obtained from

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{x}_{p+1} + \boldsymbol{\varepsilon},$$

where $\mathbf{x}_{p+1} = \mathbf{x}_1 \circ \mathbf{x}_2$ is an interaction term which is the product of the first two variables. However, we still apply the model (1): $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ to estimate the coefficients. Therefore, the true model cannot be correctly specified in this simulation study.

The first two models are similar with those in the random lasso paper by Wang et al. (2011), which are designed to investigate the performance of the existing methods for the data with complicated correlation structures. The last misspecified model is taken from Lv and Liu (2014).

To compute the solution paths for lasso and adaptive lasso, we utilized the **R** package *lassoshooting*, and while the initial weight of the adaptive lasso is obtained by the ridge estimator with a small shrinkage factor ($\lambda = 0.01$). For SCAD and MCP, **R** package *plus* is implemented to obtain the solution paths.

For each penalty, we need to initially select a grid of tuning parameters λ . For the lasso and the adaptive lasso, we select $K = 100$ values by (??). For the ridge, we apply the same grid as the lasso. For the SCAD and the MCP, the grid of tuning parameters are generated automatically by the *plus* function.

In the SPSP algorithm, the selection of the constant R is data-adaptive. Generally, for each penalty, we obtain an estimator with a small shrinkage factor ($\lambda = 0.01$), and then compute the adjacent distances of the sorted absolute values of the coefficients. Thus, the

constant R can be computed as

$$R = \frac{\text{Maximal adjacent distance}}{\text{Second maximal adjacent distance}}. \quad (8)$$

Considering the criteria for selecting the tuning parameter λ , we apply the two-fold CV to reduce the computational cost. And for the GCV, the AIC, the BIC, the EBIC, the formulas are given as follows,

$$\begin{aligned} \lambda_{GCV} &= \underset{\lambda_k}{\operatorname{argmin}} \left(\frac{SSE(\lambda_k)}{n(1 - \hat{s}_k/n)^2} \right), \\ \lambda_{AIC} &= \underset{\lambda_k}{\operatorname{argmin}} (n \log(SSE(\lambda_k)) + 2\hat{s}_k), \\ \lambda_{BIC} &= \underset{\lambda_k}{\operatorname{argmin}} (n \log(SSE(\lambda_k)) + \hat{s}_k \log n), \\ \lambda_{EBIC} &= \underset{\lambda_k}{\operatorname{argmin}} (n \log(SSE(\lambda_k)) + \hat{s}_k (2 \log p + \log n)), \end{aligned}$$

where $SSE(\lambda_k) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_k\|_2^2$, \hat{s}_k is the number of the nonzero variables at the tuning parameter λ_k , i.e., $\hat{s}_k = \sum_{j=1}^p I(\hat{\beta}_{k,j} \neq 0)$.

For each model, we record the following measures: FP (False Positive, the number of zero variables incorrectly identified as nonzero), FN (False Negative, the number of nonzero variables incorrectly identified as zero), and ME (Model Error = $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \hat{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) / \sigma^2$). Note that in each simulation, the true model is known, therefore we also record the result of the “oracle” selection in a sense that we select the “best” tuning parameter which minimizes the number of the incorrect selections. We report the mean and the standard error of all these values over $B = 100$ replicates for these six models. We also tried $B = 200$ and the results are similar.

4.1 The Main Results

The results for these models are summarized in Table 3-6. From Table 3, we can observe that the CV, the GCV, the AIC and the BIC on all the penalties tend to select too many variables in the model, which lead to large FP values. For the EBIC, the FP value is almost 0 but the FN value is large. It indicates that this criterion excludes most of the informative variables for producing an over-sparse model. It is also seen that the SPSP on the lasso, the adaptive lasso, the SCAD and the MCP all perform well in excluding the irrelevant variables in the model (small FP values). Meanwhile, the SPSP on all the penalties exclude almost half of the relevant variables in the model, which is better than the results of the EBIC. Finally, we also notice that the model errors of the SPSP are smaller than the other selection criteria for all the penalties.

Table 4-6 show the results for high dimensional models. In Table 4, we can see that the FN values of all the approaches are close on the same penalty; however, the SPSP algorithm on the penalty has much smaller FP values than the other selection criteria. It indicates that the proposed algorithm can remarkably reduce the number of the irrelevant variables for this high dimensional example. When the noise level increases from 1 to 3, the SPSP Algorithm still performs well in reducing the FP values compared with the other criteria.

Table 5 presents the simulation results for Model 3. Note, in the model, the true models are highly sparse and the relevant variables and the irrelevant variables are all correlated. We observe that the SPSP procedure can remarkably improves the selection accuracy by selecting fewer irrelevant variables on the existing penalties in general. From these two tables, we also notice that the adaptive lasso with the EBIC also has a competitive performance in selection due to the over-sparsity patterns of the true models but the model errors are much larger than those of the SPSP approaches especially when $p = 1000$.

The results of Model 4 are shown in Table 6. We observe that when models are misspecified, the SPSP algorithm on all the four penalties all have a good performance in both selection and estimation accuracy. Note that the adaptive lasso with the EBIC, the SCAD with CV, the MCP with CV all perform similarly with SPSP in terms of FP values. It is worth mentioning that the ME values are all relatively large since we divide σ^2 in computing ME and the noise level $\sigma = 0.25$ is quite small in this model.

In summary, from these simulation studies, we can see that the SPSP approach can improve the selection accuracy of the penalized least squares estimations in general, especially when high correlations among the variables are presented. It can well balance the trade-off between the model fitting and the model sparsity compared with the other criteria. Compared with the other criteria, the proposed algorithm can select fewer irrelevant variables without excluding more relevant variables for high dimensional data problems.

4.2 Sensitivity of the Constant R in SPSP

The constant R in the proposed SPSP algorithm controls the order of the magnitude in the partitioning rule. Here we examine the sensitivity of the SPSP procedure with respect to the choice of the constant R . Note that in the numerical studies we conducted, these constants are data adaptive—it is estimated from the data set, rather than given arbitrarily like the cut-off probability in the stability selection.

In particular, we select a sequence of the numbers from 1 to 10 with the increment 0.5, as the candidates of the constant R . We use each number as the constant R in the proposed SPSP algorithm on lasso for all the six models, and record the means of the false positive rates (FPR) and false negative rates (FNR) for each number of R .

The results are shown in Figure 3. We observe that the means of the FPR and FNR are relatively stable across different choices of the constant R . Note that the vertical line

in each graph represents the selection of the constant R by using (8). It illustrates that the selection results of the SPSP algorithm are not so sensitive to the choice of the constant R as long as R stays within a reasonable range.

4.3 The SPSP Algorithm on the ridge

One compelling advantage of the proposed SPSP procedure is that it can be applied for the penalties which cannot produce sparse solutions, such as the ridge penalty. Therefore we also implement the SPSP algorithm on the ridge for these examples. The results are presented in Table 7. For the convenience of comparison, we also report the ranking of the performance of the ridge among all the five penalties (the lasso, the adaptive lasso, the SCAD, the MCP and the ridge) with the SPSP procedure. For instance, if rank equals 3, it means this result of ridge with SPSP is the third smallest among these five penalties. In two of the models, the ridge returns the best results in terms of selection accuracy. Although the ME numbers of ridge are generally large compared to other penalties, the differences between the ME values of ridge and those of other penalties are relatively small.

4.4 The Comparison with the Stability Selection

The stability selection (SS) approach, proposed by Meinshausen and Bühlmann (2010), also avoids the selection issue of the tuning parameter by using the subsampling techniques. Hence, we compare the performance of the SPSP procedure with the SS approach for all the 4 models. Here we omit the case of Model 3($p = 1000$) due to the huge computational cost for stability selection. As suggested in Meinshausen and Bühlmann (2010), we evaluate the selection probabilities over 100 subsamples and choose the threshold value as 0.9. The results of the SS algorithm can be found in Table 8 and Table 9. It can be seen that

generally the SPSP algorithm has better performance than SS in terms of both selection accuracy and model errors. Particularly, the SS approach tends to exclude many relevant variables. In addition, we also notice that the computational cost of the SS algorithm is dramatically higher than that of the SPSP procedure since its process involves resampling and solution paths for all the 100 subsamples to be computed.

4.5 SPSP for Gaussian Graphical Modeling

Besides feature selection for linear model, the SPSP procedure introduced here can be widely applied for selection problems under the framework of penalized likelihood estimation. Here we simply present a simulation study to illustrate the performance of the SPSP algorithm in the Gaussian graphical model.

The data is simulated from $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, where the inverse of the covariance matrix is set as $(\mathbf{\Sigma}^{-1})_{j,j} = 1$, $(\mathbf{\Sigma}^{-1})_{j,j+1} = 0.5$, $(\mathbf{\Sigma}^{-1})_{j+1,j} = 0.5$, $j = 1, \dots, p/4$, and zero otherwise. We set $p = 100$ and $n = 50$ in the simulation. This example is a AR(1) model, which has been used by Friedman et al. (2008) and Yuan and Lin (2007) for the numerical study of the graphical lasso.

We compare the performance of the proposed SPSP algorithm with the graphical models selected by BIC and the EBIC. Here the details about using BIC and the EBIC to choose the tuning parameter in the graphical models are described in Foygel and Drton (2010). We apply the **R** package *glasso* to solve the graphical lasso estimators and apply the package *qgraph* to select the graphical lasso models by BIC and the EBIC. Note that the grid of the tuning parameters in the simulation is generated automatically by the function *glassopath* and all the graphs are drawn by the function *qgraph*.

We report the mean and the standard error of the number of the false positives (FP), the number of the false negatives (FN) of the SPSP algorithm, BIC and the EBIC over

100 replicates in Table 10. We observe that the BIC tends to include too many zero dependencies (high FP value) while the EBIC missed all the nonzero dependencies (high FN value) in the model. Compared with the results of these two criteria, the SPSP algorithm has a much better performance in terms of selection accuracy, which selects most of the nonzero dependencies without adding many zero dependencies in the model.

5 Data Analysis

In this section, we apply the SPSP approach on the glioblastoma gene expression data (McLendon et al., 2008)(<http://cancergenome.nih.gov/>), which aims at identifying the highly informative genes to explain the glioblastoma tumor behavior. In the analysis, we use all the censor subjects and take the logarithm of the survival time as the response variable. Finally we obtain a data with $n = 185$ subjects and $p = 930$ genes.

Similarly, we randomly divide the dataset into a training set (120 observations) and a test set (65 observations) 100 times to evaluate the average performance of the selection methods. Each time we standardize the data firstly; then we apply the SPSP approach and the other popular selection criteria on the lasso, the adaptive lasso and the ridge on the training set. We also record the number of the selected features (Nm) and the average prediction error (PE) by applying the estimations in the test set.

Table 11 shows the results of all the approaches. We can see that the GCV, the AIC, and the BIC selected many features in the model which makes the model excessively complicated to interpret. For the extended BIC, it cannot identify any features and proposes a simple average model. The SPSP algorithm on the lasso, the adaptive lasso and the ridge select a few informative features with small PE values. Here we notice that the average simple model yields the smallest prediction error, and the SPSP algorithm provides some

information with regard to identifying the informative genes in the model.

Specifically, RRAS2, PAK1 and FRAT1 are identified by the SPSP procedure of the lasso and the adaptive lasso on the whole the dataset. Some previous studies have demonstrated the strong relations between these genes and the glioblastoma tumor behavior. Throughout the experiments, Aoki et al. (2007) found out that the presence of phosphorylated PAK1 in the cytoplasm of glioblastoma cells is associated with shorter survival, which suggests that the PAK1 plays a role in the invasiveness of glioblastoma and it might be a potential target for the management of glioblastoma. Demuth et al. (2008) showed the RRAS2 is one of the candidate genes whose functions are linked to glioblastoma via technical validation. Guo et al. (2010) detected that the expression of the FRAT1 in human gliomas by immunohistochemistry, Western blot and RT-PCR and concluded that FRAT1 may be an important factor in the tumorigenesis and progression of gliomas. These studies all confirmed the selection accuracy of the proposed SPSP algorithm.

6 Discussion

We proposed a novel selection procedure for penalized likelihood approach based on the whole solution paths. By utilizing estimators over all the values of the tuning parameter, we can obtain better selection accuracy, compared to the commonly used approach of just selecting one tuning parameter using certain criterion. Moreover, our SPSP procedure also achieves selection consistency under conditions that are substantially weaker than the irrepresentable condition, which is almost necessary under the current framework. Another advantage is that we now can carry out selection with a strictly convex l_2 penalty. Although the paper mainly focuses on feature selection for linear models, the SPSP procedure can be easily applied to most selection problems with one or more tuning parameters.

With this manuscript, we hope to raise the discussion of better applying the information of the whole solution paths. For example, we can rank the importance of the features by exploring the differences between behaviors of the solutions paths of the important features and the spurious ones. It is also possible to develop inference procedure based on the whole solution paths. In addition, it might be interesting to see whether the solution paths of the important features differ from those of irrelevant ones in any other manners besides the magnitude of the estimators.

References

- Hiroshi Aoki, Tomohisa Yokoyama, Keishi Fujiwara, Ana M Tari, Raymond Sawaya, Dima Suki, Kenneth R Hess, Kenneth D Aldape, Seiji Kondo, Rakesh Kumar, et al. Phosphorylated pak1 level in the cytoplasm correlates with shorter survival time in patients with glioblastoma. *Clinical Cancer Research*, 13(22):6603–6609, 2007.
- Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.
- Peter J Bickel, Ya’acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.
- Małgorzata Bogdan, Jayanta K Ghosh, and RW Doerge. Modifying the schwarz bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, 167(2):989–999, 2004.
- Karl W Broman and Terence P Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):641–656, 2002.

- Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.
- Florentina Bunea, Alexandre Tsybakov, Marten Wegkamp, et al. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.
- Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.
- Haeran Cho and Piotr Fryzlewicz. High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 74(3):593–622, 2012.
- Tim Demuth, Jessica Rennert, Dominique Hoelzinger, Linsey Reavie, Mitsutoshi Nakada, Christian Beaudry, Satoko Nakada, Eric Anderson, Amanda Henrichs, Wendy McDonough, et al. Glioma cells on the run—the migratory transcriptome of 10 human glioma cell lines. *BMC genomics*, 9(1):54–68, 2008.
- Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
- Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.
- Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in Neural Information Processing Systems*, 2010.

- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- Geng Guo, Xinggang Mao, Peng Wang, Bolin Liu, Xiang Zhang, Xiaofan Jiang, Chengliang Zhong, Junli Huo, Ji Jin, and Yuzhen Zhuo. The expression profile of *frat1* in human gliomas. *Brain research*, 1320:152–158, 2010.
- Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, 14(4):382–401, 1999.
- Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *The Annals of statistics*, 38(4):2282–2313, 2010.
- Jinchi Lv and Jun S Liu. Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):141–167, 2014.
- Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogiannis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.
- Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.
- Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
- David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004.

- Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.
- Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.
- David Siegmund. Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika*, 91(4):785–800, 2004.
- Wei Sun, Junhui Wang, and Yixin Fang. Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research*, 14(1):3419–3440, 2013.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.
- Sara van de Geer. The deterministic lasso. In *Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich*, 2007.
- Hansheng Wang, Guodong Li, and Chih-Ling Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78, 2007.

- Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *The Annals of Applied Statistics*, 5(1):468–485, 2011.
- Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.
- Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.
- Yiyun Zhang, Runze Li, and Chih-Ling Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 2010.
- Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.
- Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Table 1: The general selection criteria in statistics and machine learning literature.

CRITERIA	METHOD	REFERENCES
CV	ADAPTIVE LASSO	ZOU (2006), JASA
	FUSED LASSO	TIBSHIRANI ET AL. (2005), JRSSB
	TLP	SHEN ET AL. (2012), JASA
GCV	LASSO	TIBSHIRANI (1996), JRSSB
	SCAD	FAN AND LI (2001), JASA
BIC	RA-LASSO	WANG ET AL. (2007), JRSSB
	LASSO-TYPE	YUAN AND LIN (2007), BIOMETRIKA
EBIC	TILTING	CHO AND FRYZLEWICZ (2012), JRSSB
	GROUP LASSO	HUANG ET AL. (2010), AOS

Table 2: The number of the selected variables, the number of false positives (FP), the number of false negatives (FN) of the criteria: CV, GCV, AIC, BIC, EBIC, Oracle. Note that the true model contains 10 nonzero variables and 30 zero variables.

	CV	GCV	AIC	BIC	EBIC	Oracle
TNS	16	35	37	35	4	4
FP	12	27	30	27	0	0
FN	6	2	3	2	6	6

Table 3: Results of Model 1 (low-dim) over 100 replicates for the lasso, the Alasso (Adaptive lasso), the SCAD, the MCP (Standard Error in the parentheses).

		SPSP	2-CV	GCV	AIC	BIC	EBIC	Oracle
lasso	FP	0.63	8.31	10.42	26.12	5.03	0.77	0.82
		(0.24)	(1.61)	(1.28)	(0.77)	(1.32)	(0.13)	(0.13)
	FN	5.11	5.58	5.79	1.69	6.27	7.03	6.95
		(0.28)	(0.39)	(0.32)	(0.27)	(0.33)	(0.13)	(0.12)
	ME	0.44	0.59	0.53	0.72	0.59	0.57	0.57
		(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
Alasso	FP	0.65	5.63	10.62	22.59	4.12	0.19	0.37
		(0.20)	(1.44)	(0.91)	(1.01)	(0.97)	(0.07)	(0.09)
	FN	4.03	6.57	4.55	1.40	6.73	8.09	7.19
		(0.29)	(0.43)	(0.38)	(0.24)	(0.33)	(0.11)	(0.18)
	ME	0.40	0.80	0.51	0.69	0.55	0.63	0.55
		(0.02)	(0.04)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)
SCAD	FP	1.09	0.86	12.21	15.03	5.07	0.00	0.24
		(0.37)	(0.28)	(0.68)	(0.61)	(0.88)	(0.00)	(0.08)
	FN	5.86	8.34	6.25	5.84	7.10	8.94	6.90
		(0.17)	(0.13)	(0.16)	(0.18)	(0.20)	(0.05)	(0.16)
	ME	0.47	0.61	0.66	0.71	0.57	0.61	0.66
		(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.04)
MCP	FP	0.96	0.61	12.04	15.38	6.92	0.00	0.97
		(0.35)	(0.24)	(0.63)	(0.66)	(0.82)	(0.00)	(0.17)
	FN	5.31	8.86	5.79	5.39	6.47	8.94	6.88
		(0.20)	(0.07)	(0.18)	(0.17)	(0.21)	(0.05)	(0.22)
	ME	0.45	0.61	0.68	0.73	0.61	0.60	0.48
		(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)

Table 4: Results of Model 2 (high-dim) over 100 replicates for the lasso, the Alasso (Adaptive lasso), the SCAD, the MCP (Standard Error in the parentheses).

$\sigma = 1$		SPSP	2-CV	GCV	AIC	BIC	EBIC	Oracle
lasso	FP	0.04	24.60	49.74	49.48	49.27	8.89	0.01
		(0.03)	(1.66)	(0.29)	(0.28)	(0.28)	(2.58)	(0.01)
	FN	2.11	1.97	1.68	1.68	1.68	2.00	2.01
		(0.06)	(0.02)	(0.08)	(0.08)	(0.08)	(0.03)	(0.01)
	ME	1.54	1.21	1.35	1.35	1.35	2.09	2.54
		(0.07)	(0.07)	(0.05)	(0.05)	(0.05)	(0.15)	(0.17)
Alasso	FP	0.02	9.20	40.47	42.95	42.86	0.1	0.08
		(0.02)	(1.18)	(0.44)	(0.19)	(0.20)	(0.05)	(0.04)
	FN	1.05	1.05	0.24	0.22	0.22	2.1	1.92
		(0.13)	(0.12)	(0.06)	(0.06)	(0.06)	(0.06)	(0.05)
	ME	1.00	1.14	1.31	1.33	1.33	1.8	1.82
		(0.09)	(0.07)	(0.05)	(0.05)	(0.05)	(0.12)	(0.11)
SCAD	FP	0.02	1.52	39.08	40.03	39.69	20.02	0.01
		(0.02)	(0.43)	(1.06)	(0.98)	(1.03)	(3.28)	(0.01)
	FN	3.42	3.89	3.68	3.68	3.68	3.81	2.46
		(0.11)	(0.05)	(0.09)	(0.09)	(0.09)	(0.08)	(0.09)
	ME	2.27	2.48	1.36	1.36	1.36	2.07	6.40
		(0.09)	(0.11)	(0.05)	(0.05)	(0.05)	(0.13)	(0.38)
MCP	FP	0.01	0.85	46.47	46.54	46.48	42.74	0.03
		(0.01)	(0.24)	(0.67)	(0.54)	(0.55)	(2.03)	(0.02)
	FN	3.59	3.96	3.68	3.76	3.78	3.80	3.68
		(0.10)	(0.03)	(0.10)	(0.09)	(0.08)	(0.08)	(0.08)
	ME	2.48	2.76	1.35	1.36	1.36	1.50	2.62
		(0.10)	(0.11)	(0.55)	(0.05)	(0.05)	(0.08)	(0.10)
$\sigma = 3$		SPSP	2-CV	GCV	AIC	BIC	EBIC	Oracle
lasso	FP	0.87	33.45	51.57	51.39	51.37	38.91	1.75
		(0.28)	(0.72)	(0.31)	(0.31)	(0.31)	(2.81)	(0.20)
	FN	2.96	2.25	1.96	1.95	1.95	2.13	2.31
		(0.13)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)
	ME	0.39	0.66	1.02	1.02	1.02	0.94	0.51
		(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
Alasso	FP	1.36	22.84	43.06	44.42	44.19	0.19	0.30
		(0.47)	(1.15)	(0.31)	(0.19)	(0.20)	(0.07)	(0.07)
	FN	2.61	1.87	1.21	1.11	1.10	3.02	2.51
		(0.15)	(0.13)	(0.13)	(0.12)	(0.12)	(0.10)	(0.10)
	ME	0.39	0.58	1.01	1.02	1.02	0.53	0.44
		(0.05)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)
SCAD	FP	1.37	3.36	39.23	40.04	39.74	21.81	0.02
		(0.57)	(0.54)	(1.12)	(1.06)	(1.09)	(3.28)	(0.02)
	FN	3.66	3.91	3.94	3.94	3.94	3.99	2.76
		(0.08)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.10)
	ME	0.40	0.42	1.02	1.02	1.02	0.74	0.88
		(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.07)	(0.05)
MCP	FP	1.69	1.90	45.37	45.93	45.81	39.64	0.02
		(0.70)	(0.43)	(0.90)	(0.79)	(0.85)	(2.56)	(0.02)
	FN	3.95	4.02	3.93	3.94	3.94	3.96	3.96
		(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
	ME	0.45	0.46	1.02	1.02	1.02	0.94	0.46
		(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)

Table 5: Results of Model 3 (sparse model) over 100 replicates for the lasso, the Alasso (Adaptive lasso), the SCAD, the MCP (Standard Error in the parentheses).

$p = 100$									$p = 1000$								
		SPSP	2-CV	GCV	AIC	BIC	EBIC	Oracle			SPSP	2-CV	GCV	AIC	BIC	EBIC	Oracle
lasso	FP	0.23	22.04	49.82	49.87	49.87	11.72	0.37	lasso	FP	0.04	41.55	56.85	56.56	56.03	0.11	0.14
		(0.08)	(1.49)	(0.29)	(0.26)	(0.26)	(2.88)	(0.08)			(0.03)	(2.29)	(0.42)	(0.41)	(0.39)	(0.04)	(0.05)
	FN	0.34	0.00	0.02	0.02	0.02	0.01	0.00		FN	0.75	0.00	0.00	0.00	0.00	0.69	0.07
		(0.09)	(0.00)	(0.02)	(0.02)	(0.02)	(0.01)	(0.00)			(0.12)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.04)
	ME	0.57	0.43	1.02	1.02	1.02	0.62	0.50		ME	0.61	0.82	1.05	1.05	1.05	1.50	1.09
(0.06)		(0.04)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)	(0.09)	(0.05)		(0.03)	(0.03)	(0.03)	(0.17)	(0.08)		
Alasso	FP	0.86	13.28	42.95	45.55	45.46	0.16	0.10	Alasso	FP	0.13	20.08	41.70	44.78	44.78	0.11	0.09
		(0.25)	(1.28)	(0.78)	(0.15)	(0.15)	(0.06)	(0.04)			(0.06)	(1.92)	(0.87)	(0.29)	(0.29)	(0.04)	(0.04)
	FN	0.21	0.00	0.01	0.00	0.01	0.10	0.00		FN	0.57	0.01	0.00	0.00	0.00	0.51	0.12
		(0.07)	(0.00)	(0.01)	(0.00)	(0.01)	(0.04)	(0.00)			(0.11)	(0.01)	(0.00)	(0.00)	(0.00)	(0.12)	(0.05)
	ME	0.23	0.31	0.98	1.01	1.01	0.23	0.24		ME	0.49	0.53	0.98	1.02	1.02	0.79	0.61
(0.04)		(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.08)	(0.04)		(0.04)	(0.03)	(0.03)	(0.13)	(0.06)		
SCAD	FP	2.85	2.84	31.98	33.41	32.58	9.68	0.11	SCAD	FP	1.95	6.20	29.56	31.61	30.70	15.45	0.07
		(0.79)	(0.45)	(1.23)	(1.14)	(1.23)	(2.59)	(0.04)			(0.43)	(0.77)	(0.60)	(0.70)	(0.69)	(2.19)	(0.04)
	FN	0.32	0.24	0.11	0.11	0.11	0.37	0.13		FN	0.50	0.27	0.37	0.37	0.37	0.59	0.13
		(0.08)	(0.06)	(0.04)	(0.04)	(0.04)	(0.07)	(0.05)			(0.09)	(0.06)	(0.07)	(0.07)	(0.07)	(0.12)	(0.05)
	ME	0.42	0.35	1.00	1.00	1.00	0.53	0.52		ME	0.58	0.57	1.04	1.04	1.04	1.18	0.95
(0.06)		(0.05)	(0.03)	(0.03)	(0.03)	(0.05)	(0.06)	(0.06)	(0.07)		(0.03)	(0.03)	(0.03)	(0.15)	(0.09)		
MCP	FP	1.38	1.47	42.56	43.94	43.84	29.13	0.16	MCP	FP	0.66	2.25	46.06	46.07	46.02	44.81	0.03
		(0.43)	(0.41)	(0.94)	(0.78)	(0.80)	(3.24)	(0.05)			(0.25)	(0.46)	(0.54)	(0.49)	(0.53)	(1.31)	(0.02)
	FN	0.37	0.40	0.12	0.12	0.12	0.25	0.13		FN	0.69	0.44	0.30	0.30	0.30	0.30	0.30
		(0.08)	(0.07)	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)			(0.10)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
	ME	0.45	0.46	1.02	1.02	1.02	0.94	0.46		ME	0.58	0.58	1.04	1.04	1.04	1.03	0.48
(0.06)		(0.06)	(0.03)	(0.03)	(0.03)	(0.06)	(0.04)	(0.07)	(0.06)		(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	

Table 6: Results of Model 4 (misspecified model) over 100 replicates for the lasso, the Alasso (Adaptive lasso), the SCAD, the MCP (Standard Error in the parentheses).

		SPSP	2-CV	GCV	AIC	BIC	EBIC	Oracle
lasso	FP	0.39	27.20	44.72	45.47	45.47	23.35	0.66
		(0.13)	(1.65)	(0.34)	(0.17)	(0.17)	(3.14)	(0.13)
	FN	0.95	0.06	0.03	0.03	0.03	0.83	0.38
		(0.16)	(0.06)	(0.02)	(0.02)	(0.02)	(0.21)	(0.09)
	ME	14.66	13.71	18.77	18.88	18.88	27.21	28.91
		(1.97)	(1.16)	(1.04)	(1.04)	(1.04)	(2.75)	(1.81)
Alasso	FP	0.67	19.09	36.91	42.18	42.15	0.30	0.37
		(0.17)	(1.59)	(1.14)	(0.25)	(0.25)	(0.08)	(0.08)
	FN	0.77	0.20	0.11	0.08	0.08	1.55	0.36
		(0.14)	(0.08)	(0.04)	(0.04)	(0.04)	(0.26)	(0.08)
	ME	12.94	12.86	17.21	18.29	18.29	26.99	16.14
		(1.78)	(1.32)	(1.03)	(1.02)	(1.02)	(3.70)	(1.28)
SCAD	FP	5.28	4.00	37.86	38.67	38.47	25.71	0.45
		(1.17)	(0.73)	(1.07)	(0.97)	(1.01)	(3.01)	(0.10)
	FN	0.48	0.44	0.14	0.14	0.14	0.58	0.16
		(0.14)	(0.15)	(0.06)	(0.06)	(0.06)	(0.19)	(0.06)
	ME	14.77	18.13	18.82	18.84	18.83	22.65	16.68
		(2.13)	(2.36)	(1.03)	(1.03)	(1.03)	(3.33)	(1.70)
MCP	FP	2.50	1.96	44.00	44.30	44.11	40.96	0.16
		(0.75)	(0.46)	(0.58)	(0.43)	(0.56)	(1.84)	(0.07)
	FN	0.57	0.65	0.11	0.11	0.11	0.23	0.15
		(0.15)	(0.16)	(0.05)	(0.05)	(0.05)	(0.11)	(0.07)
	ME	13.96	16.97	18.90	18.93	18.90	19.37	10.69
		(2.23)	(2.46)	(1.04)	(1.04)	(1.04)	(1.61)	(1.16)

Table 7: Simulation results of the SPSP approach on the ridge (Standard Error in parentheses), ranking among all the five penalties in the third row.

		M1	M2($\sigma = 1$)	M2($\sigma = 3$)	M3($p = 100$)	M3($p = 1000$)	M4
ridge	FP	4.32	0.52	2.33	0.36	0.18	0.37
		(0.81)	(0.16)	(0.61)	(0.11)	(0.07)	(0.11)
		5	5	5	2	3	1
	FN	3.38	0.48	1.50	0.61	0.96	2.19
		(0.32)	(0.10)	(0.21)	(0.10)	(0.10)	(0.18)
		1	1	1	5	5	5
	ME	0.49	0.77	0.53	0.55	0.86	31.33
		(0.03)	(0.08)	(0.08)	(0.08)	(0.08)	(2.77)
		5	1	5	4	5	5

Table 8: Results of the SS algorithm and the SPSP on the lasso.

		M1	M2($\sigma = 1$)	M2($\sigma = 3$)	M3($p = 100$)	M4
SS	FP	0.35	0.01	0.01	0.10	0.00
		(0.64)	(0.10)	(0.10)	(0.32)	(0.00)
	FN	9.47	2.37	4.79	0.30	2.30
		(0.56)	(0.54)	(0.73)	(0.48)	(0.82)
	ME	1.93	1.68	1.64	0.26	28.02
		(1.31)	(0.57)	(0.97)	(0.34)	(9.93)
SPSP	FP	0.63	0.87	0.01	0.23	0.39
		(0.24)	(0.03)	(0.28)	(0.08)	(0.13)
	FN	5.11	2.11	2.96	0.34	0.95
		(0.28)	(0.54)	(0.09)	(0.48)	(0.16)
	ME	0.44	1.54	0.39	0.57	14.66
		(0.02)	(0.07)	(0.04)	0.06	(1.97)

Table 9: The average time for computing the SS and the SPSP estimators (in seconds).

	M1	M2($\sigma = 1$)	M2($\sigma = 3$)	M3($p = 100$)	M4
SS	41.23s	88.79s	121.16s	199.29s	162.39s
SPSP	0.44s	1.55s	0.39s	3.89s	2.90s

Table 10: The mean of FP, FN values of the SPSP algorithm, BIC, and the EBIC over 100 replicates (Standard Error in the parentheses). The true model has 25 nonzero dependencies and 4925 zero dependencies.

	SPSP	BIC	EBIC
FP	19.31	116.56	0
	(2.48)	(3.2)	(0)
FN	2.50	0	25
	(0.80)	(0.0)	(0)

Table 11: Results of glioblastoma gene expression analysis.

lasso	SPSP	GCV	AIC	BIC	EBIC
Nm	12.81 (0.989)	139.65 (0.433)	140.92 (0.467)	140.65 (0.438)	0.00 (0.000)
PE	1.25 (0.025)	1.62 (0.024)	1.62 (0.024)	1.62 (0.024)	1.00 (0.02)
Alasso	SPSP	GCV	AIC	BIC	EBIC
Nm	8.58 (0.561)	70.88 (1.554)	71.53 (1.538)	44.96 (3.520)	0.04 (0.020)
PE	1.22 (0.023)	1.31 (0.023)	1.31 (0.024)	1.22 (0.025)	1.00 (0.019)

ridge	SPSP
Nm	3.29 (0.330)
PE	1.08 (0.021)

Figure 1: Left: The lasso solution paths of the simulated example. The dashed lines are the paths of the 10 nonzero variables while the black lines are the paths of the 30 zero variables. The vertical lines represent the tuning parameter selected by the criteria. Right: The lasso solution paths of the nonzero variables “1”, “2” and the zero variable “3”.

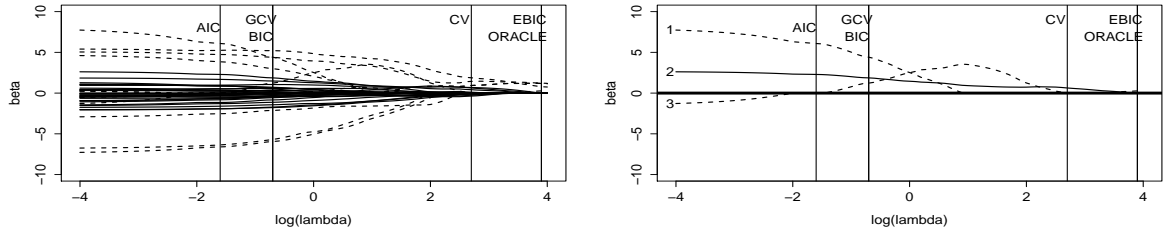


Figure 2: Left: Partitions on the lasso solution paths of the same simulated example. Right: Partitions on the lasso solution paths of nonzero coefficients “1”, “3” and zero coefficient “2”.

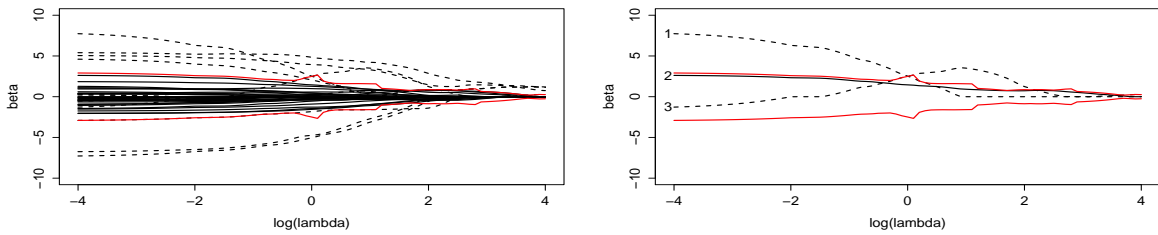
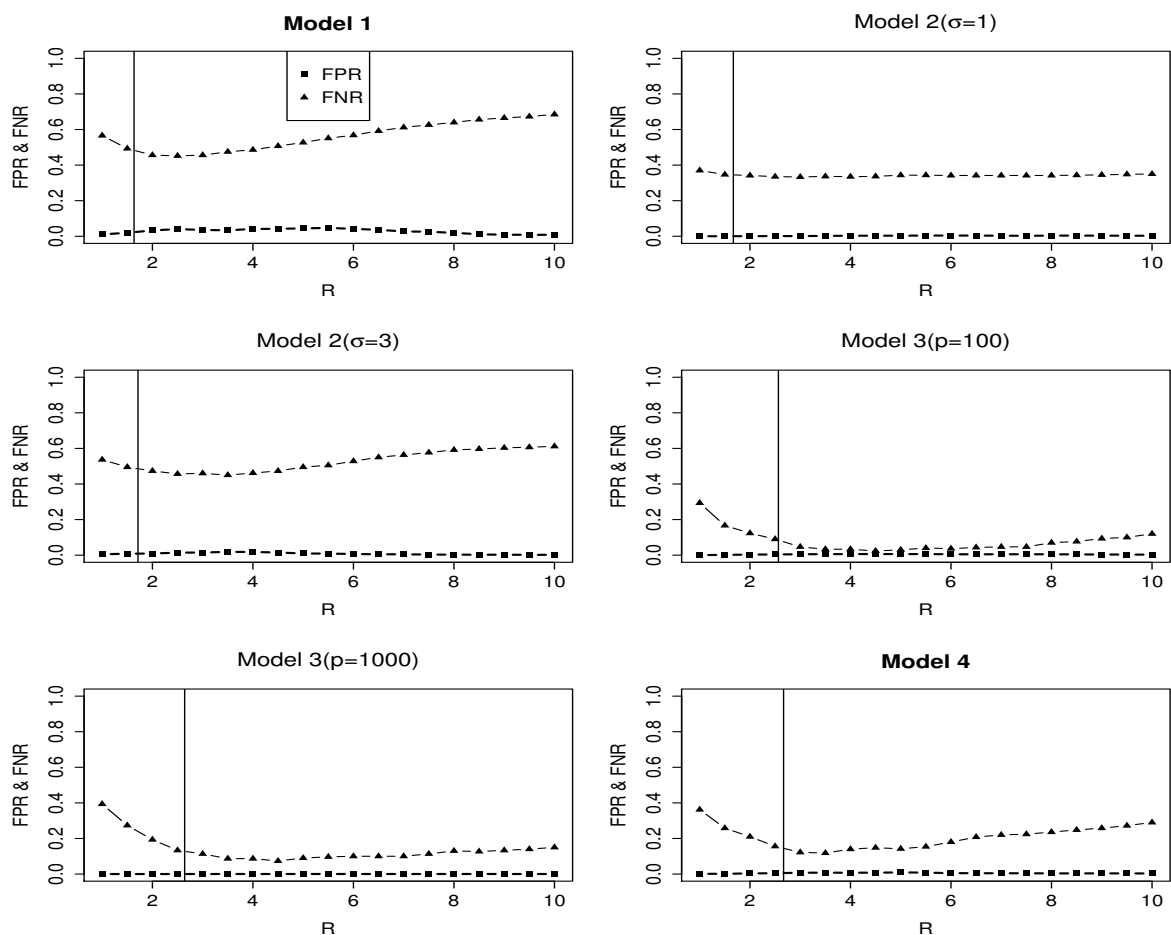


Figure 3: The mean of the FPR and FNR over 100 replicates over different choices of the constant R in SPSP on the Lasso. The vertical lines in the graphs are the average values of the selected values by (8).



Supplementary File for Selection by Partitioning the Solution Paths: Technical Proofs

Yang Liu

Fred Hutchinson Cancer Research Center

and

Peng Wang

Department of Operations, Business Analytics and Information Systems
University of Cincinnati

June 22, 2016

Lemma 1. *Suppose the compatibility condition holds. Let $\lambda_0 = 2\sigma\sqrt{\frac{t^2+2\log p}{n}}$ for any $t > 0$, then for $\lambda \geq 2\lambda_0$, with probability at least $1 - 2e^{-t^2/2}$, we have*

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{4\lambda^2 s}{\phi^2}. \quad (1)$$

Proof. The proof follows from Lemma 6.2 and Theorem 6.1 in Bühlmann and van de Geer (2011). \square

It follows apparently from Lemma 1 that we can bound the bias term by

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{4\lambda s}{\phi^2}.$$

Let $\delta_\lambda = \frac{4\lambda s}{\phi^2}$ and $\delta_0 = \delta_{\lambda_0}$, where λ_0 is defined in Lemma 1.

We first sort the absolute values true non-zero coefficients in ascending order to get $|\beta^*|_{(1)}, \dots, |\beta^*|_{(s)}$, and define the true adjacent distance as $D_0 = 0, D_1 = |\beta^*|_{(1)}, D_2 =$

$|\beta^*|_{(2)} - |\beta^*|_{(1)}, \dots, D_s = |\beta^*|_{(s)} - |\beta^*|_{(s-1)}$. Let $C_0 = \sqrt{\frac{D_{\max} + \delta_0}{\delta_0}} - 1$, $D_{\max} = \max_{1 \leq i \leq p} \{D_i\}$ and

$$C = \frac{D_{\max}}{\min\{|\beta_i^*| : |\beta_i^*| > (2 + C_0)\delta_0\}}.$$

Moreover, let

$$C_{\text{under}}^i = \frac{D_i}{\max\{D_{i'} : i' < i\}},$$

for $i = 2, \dots, s$ and $C_{\text{under}}^1 = \infty$, and

$$C_{\text{upper}}^i = \frac{D_i}{\max\{D_{i'} : i' > i\}},$$

for $i = 1, \dots, s-1$ and $C_{\text{upper}}^s = \infty$. Further define $\hat{D}(\Theta)$ as the largest adjacent distance of $\hat{\beta}_i$ for all $i \in \Theta$, where $\hat{\beta}_i$ are obtained with tuning parameter λ , i.e. $\hat{D}(\Theta) = \max_{j \in \Theta} \min_{j' \in \Theta} \left| |\hat{\beta}_j| - |\hat{\beta}_{j'}| \right|$, and $\hat{D}(\Theta_1, \Theta_2) = \min_{j \in \Theta_1, j' \in \Theta_2} \left| |\hat{\beta}_j| - |\hat{\beta}_{j'}| \right|$. In addition, let $R = 1 + C$.

Theorem 1. *Let $i_\lambda = \min\{i : C_{\text{under}}^i \geq R, C_{\text{upper}}^i \geq \frac{1}{C}, D_i > (1 - \frac{R}{C_{\text{under}}^i})^{-1}(1 + R)\delta_\lambda\}$ and $S_\lambda = \{j : |\beta_j^*| \geq |\beta^*|_{(i_\lambda)}\}$. Under the compatibility condition, if $\lambda > 2\lambda_0$, the following inequalities hold for lasso estimator with probability at least $1 - 2e^{-t^2/2}$,*

$$\frac{\hat{D}(S_\lambda, S_\lambda^c)}{\hat{D}(S_\lambda^c)} > R, \tag{2}$$

$$\frac{\hat{D}(S_\lambda)}{\hat{D}(S_\lambda, S_\lambda^c)} \leq R. \tag{3}$$

Proof. By Lemma 1, with probability at least $1 - 2e^{-t^2/2}$, we have

$$\hat{D}(S_\lambda^c) \leq \max\{D_{i'} : i' < i_\lambda\} + \delta_\lambda$$

and similarly for any $j_1 \in S_\lambda, j_2 \in S_\lambda^c$,

$$\begin{aligned} \left| |\hat{\beta}_{j_1}| - |\hat{\beta}_{j_2}| \right| &\geq |\beta_{j_1}^*| - \left| |\hat{\beta}_{j_1}| - |\beta_{j_1}^*| \right| - |\hat{\beta}_{j_2}| \\ &\geq D_{i_\lambda} - \delta_\lambda \end{aligned}$$

Then with probability at least $1 - 2e^{-t^2/2}$,

$$\frac{\hat{D}(S_\lambda, S_\lambda^c)}{\hat{D}(S_\lambda^c)} \geq \frac{D_{i_\lambda} - 2\delta_\lambda}{\max\{D_{i'} : i' < i_\lambda\} + \delta_\lambda} > R.$$

The above inequality follows immediately from $C_{\text{under}}^{i_\lambda} > R$ and $D_{i_\lambda} > (1 - \frac{R}{C_{\text{under}}^{i_\lambda}})^{-1}(1 + R)\delta_\lambda$.

To prove (3), we have

$$\hat{D}(S_\lambda) \leq D_{\max} + \delta_\lambda$$

Then with probability at least $1 - 2e^{-t^2/2}$,

$$\begin{aligned} \frac{\hat{D}(S_\lambda)}{\hat{D}(S_\lambda, S_\lambda^c)} &\leq \frac{D_{\max} + \delta_\lambda}{D_{i_\lambda} - \delta_\lambda}, \\ &\leq C + \frac{(C+1)\delta_\lambda}{D_{i_\lambda} - 2\delta_\lambda} \\ &\leq C + \frac{1+C}{R} = R. \end{aligned}$$

The last inequality follows from the fact that $R = 1 + C$. □

Theorem 2. Let $i_{2\lambda_0} = \min\{i : C_{\text{under}}^i \geq R, C_{\text{upper}}^i \geq \frac{1}{C}, D_i > (1 - \frac{R}{C_{\text{under}}^i})^{-1}(1 + R)2\delta_{\lambda_0}\}$ and $S_{2\lambda_0} = \{j : |\beta_j^*| \geq |\beta_{(i_{2\lambda_0})}^*|\}$. Under the compatibility condition, the SPSP procedure \hat{S} over $\lambda \in [2\lambda_0, \frac{\phi^2 D_{\max}}{4s(1+R)})$ recovers $S_{2\lambda_0}$ with probability at least $1 - 2e^{-t^2}$,

$$P(\hat{S} = S_{2\lambda_0}) > 1 - 2e^{-t^2}.$$

In particular, when $\min_{j \in S} |\beta_j^*| > (2 + R)2\delta_0$,

$$P(\hat{S} = S) > 1 - 2e^{-t^2}.$$

Proof. A sufficient condition for $\hat{S}_\lambda \cap S^C = \emptyset$ is

$$D_{\max} > R\delta_\lambda.$$

Then the theorem follows immediately from Theorem 1 and the fact that $\hat{S} = \cup_{\lambda} \hat{S}_{\lambda}$. \square

Lemma 2. *The irrerepresentable condition implies the identifiability condition.*

Proof.

$$\begin{aligned} & \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S \hat{\boldsymbol{\beta}}_S - \mathbf{X}_{S^c} \hat{\boldsymbol{\beta}}_{S^c}\|^2 \\ \geq & \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S \hat{\boldsymbol{\beta}}_S - \mathbf{X}_S \text{diag}(\text{sign}(\boldsymbol{\beta}_S^*)) \text{sign}(\boldsymbol{\beta}_S^*) (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{X}_{S^c} \hat{\boldsymbol{\beta}}_{S^c}\|^2 \end{aligned}$$

By the irrerepresentable condition, there is a $\eta > 0$ such that

$$\|\text{sign}(\boldsymbol{\beta}_S^*) (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{X}_{S^c}\|_{\infty} \leq 1 - \eta,$$

from which it follows immediately that $\|\text{diag}(\text{sign}(\boldsymbol{\beta}_S^*)) \text{sign}(\boldsymbol{\beta}_S^*) (\mathbf{X}_S^T \mathbf{X}_S)^{-1} \mathbf{X}_S^T \mathbf{X}_{S^c} \hat{\boldsymbol{\beta}}_{S^c}\|_1 \leq \|\hat{\boldsymbol{\beta}}_{S^c}\|_1 (1 - \eta)$. Therefore, the identifiability condition holds if the irrerepresentable condition holds. \square

Theorem 3. *Under $WIC(k, \kappa)$, the following inequality holds for the lasso solution $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\beta}}_{S^c})$ with $\lambda > \lambda_0(\frac{2+2k-k\eta}{k\eta} + \kappa)$ with probability at least $1 - 2e^{-t^2}$,*

$$\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 \leq k \|\hat{\boldsymbol{\beta}}_S\|_1.$$

Proof. Since $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\beta}}_{S^c})$ is the lasso solution, then for

$$\tilde{\boldsymbol{\beta}}_S = \arg \min_{\|\boldsymbol{\beta}_S\|_1 \leq \|\hat{\boldsymbol{\beta}}_S\|_1 + (1-\eta)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S \boldsymbol{\beta}_S\|^2,$$

we have

$$\frac{1}{n} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S \tilde{\boldsymbol{\beta}}_S\| + \lambda \|\tilde{\boldsymbol{\beta}}_S\|_1 + \frac{1}{n} 2\varepsilon^T \mathbf{X}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}) \geq \frac{1}{n} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S \hat{\boldsymbol{\beta}}_S - \mathbf{X}_{S^c} \hat{\boldsymbol{\beta}}_{S^c}\| + \lambda \|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{n} 2\varepsilon^T \mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}).$$

If the following inequality holds when $\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 > k \|\hat{\boldsymbol{\beta}}_S\|_1$

$$\lambda \|\hat{\boldsymbol{\beta}}\|_1 - \lambda \|\tilde{\boldsymbol{\beta}}_S\|_1 + \frac{1}{n} 2\varepsilon^T \mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \geq 0.$$

Then it follows from the identifiability condition that

$$\frac{1}{n} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S \tilde{\boldsymbol{\beta}}_S\| + \lambda \|\tilde{\boldsymbol{\beta}}_S\|_1 + 2\varepsilon^T \mathbf{X}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}) \leq \frac{1}{n} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S \hat{\boldsymbol{\beta}}_S - \mathbf{X}_{S^c} \hat{\boldsymbol{\beta}}_{S^c}\| + \lambda \|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{n} 2\varepsilon^T \mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}),$$

therefore either $\hat{\beta}_{S^c} = 0$ or $\|\hat{\beta}_{S^c}\|_1 \leq k\|\hat{\beta}_S\|_1$.

Because we have concluded $P(\|\frac{1}{n}2\varepsilon^T\mathbf{X}\|_\infty < \lambda_0) > 1 - 2e^{-t^2}$, when $\|\hat{\beta}_{S^c}\|_1 > k\|\hat{\beta}_S\|_1$,

$$\begin{aligned}
\lambda\|\hat{\beta}\|_1 - \lambda\|\tilde{\beta}_S\|_1 + \frac{1}{n}2\varepsilon^T\mathbf{X}(\tilde{\beta} - \hat{\beta}) &\geq \lambda\|\hat{\beta}\|_1 - \lambda\|\tilde{\beta}_S\|_1 - \lambda_0\|\tilde{\beta} - \hat{\beta}\|_1 \\
&\geq \lambda\eta\|\hat{\beta}_{S^c}\|_1 - \lambda_0\|\tilde{\beta}_S\|_1 - \lambda_0\|\hat{\beta}\|_1 \\
&\geq \lambda\eta\|\hat{\beta}_{S^c}\|_1 - \lambda_0(\frac{1}{k} + 1 - \eta)\|\hat{\beta}_{S^c}\|_1 - \lambda_0(\frac{1}{k} + 1)\|\hat{\beta}_{S^c}\|_1 \\
&= \{\lambda\eta - \lambda_0(\frac{2}{k} + 2 - \eta)\}\|\hat{\beta}_{S^c}\|_1 \\
&> 0.
\end{aligned}$$

The last inequality follows from $\lambda > \lambda_0 \frac{2+2k-k\eta}{k\eta}$.

□

Theorem 4. Under the weak identifiability condition with $k = \frac{2}{2s+Rs(s+1)}$,

$$P(\hat{S}_\lambda \subset S) \geq 1 - 2e^{-t^2},$$

for $\lambda > \lambda_0(\frac{2+2k-k\eta}{k\eta} + \kappa)$.

Proof. Denote $\hat{\beta}_{S^c}^{\max} = \max\{|\hat{\beta}_j| : j \in S^c\}$, and sort the absolute values in $\{|\hat{\beta}_j| : j \in S, |\hat{\beta}_j| \geq \hat{\beta}_{S^c}^{\max}\}$ in ascending order to get $\hat{\beta}_{(1)}^u \leq \hat{\beta}_{(2)}^u \leq \dots \leq \hat{\beta}_{(d)}^u$, where $d \leq s$ is the cardinality of the set $\{|\hat{\beta}_j| : j \in S, |\hat{\beta}_j| \geq \hat{\beta}_{S^c}^{\max}\}$. Let $\Delta_1 = \hat{\beta}_{(1)}^u - \hat{\beta}_{S^c}^{\max}$ and $\Delta_i = \hat{\beta}_{(i)}^u - \hat{\beta}_{(i-1)}^u$, $i = 2, \dots, d$, then $\|\hat{\beta}_S\|_1 \leq s\hat{\beta}_{S^c}^{\max} + \sum_{i=1}^d i\Delta_i$. Therefore by Theorem 3

$$\hat{\beta}_{S^c}^{\max} \leq \|\hat{\beta}_{S^c}\|_1 \leq ks\hat{\beta}_{S^c}^{\max} + k \sum_{i=1}^d i\Delta_i \leq ks\hat{\beta}_{S^c}^{\max} + k \frac{s(s+1)}{2} \Delta_{\max},$$

where Δ_{\max} is the maximum value of $\Delta_1, \dots, \Delta_d$. It follows when $k = \frac{2}{2s+Rs(s+1)}$ that

$$\hat{\beta}_{S^c}^{\max} \leq \frac{ks(s+1)}{2(1-ks)} \Delta_{\max} = \frac{1}{R} \Delta_{\max}.$$

□

Theorem 5. *Under the compatibility condition and the weak identifiability condition with $k = \frac{2}{2s+Rs(s+1)}$, suppose*

$$D_{\max} > \lambda_0 \frac{4s(1+R)}{\phi^2} \left\{ \frac{Rs^2 + (2+R)S + 2}{\eta} - 1 + \kappa \right\},$$

then the SPSP procedure over $\lambda \in [2\lambda_0, \infty)$ identifies $S_{2\lambda_0} = \{j : |\beta_j^| > (1+R)2\delta_0\}$ with probability at least $1 - 2e^{-t^2}$, i.e.*

$$P(\hat{S} = S_{2\lambda_0}) > 1 - 2e^{-t^2}.$$

Proof. The theorem follows directly from Theorem 2 and Theorem 4. □

References

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.